# بين خصوصية الفرد وتقنيات المعلوماتية للكشف عن الإرهاب

قاسم محمد دنش[1]

**تُحسبُ إشكالية خصوصية المواطن في لبنان وماهية المعلومات التي يمكن لأجهزة الأمن الإطلاع عليها، من أكثر المواضيع سجالًا بين الساسة اليوم، خصوصًا مع ربط الأجهزة الأمنية بالفرق السياسية. هذه الإشكالية التي أودت إلى سجالات واسعة بين اللبنانيين بعد العام 2005، على خلفية كشف أو حجب "داتا الإتصالات" لفرع المعلومات، دفعت البعض إلى التصريح العلني "انه اذا كان هناك خيار في كشف خصوصية المواطنين ومنع الجريمة، نحن حتمًا ننحاز الى منع حصول الجرائم ولو أدى الى كشف بعض خصوصيات المواطنين".**

بين هذه الإشكالية وبين قدرات السلطات الأمنية والقضائية التي تستخدم وسائل باتت تقليدية، فلا بد من تطوير آليات التحقيقات القضائية، وكذلك الترتيبات الأمنية، بما يتناسب مع هذا الزمان الذي تعدّ فيه التكنولوجيا محورًا لا يتجزأ، إن بالأعمال الإرهابية والجرائم المنظمة، وإن بآليات الكشف عن هذه الجرائم وتتبعها، والكشف عنها قبل حصولها أيضًا. فهل من الممكن الاستفادة من مزايا عصر ثورة المعلوماتية لتحقيق نتائج مرجوة من تحليل لمعلومات وبيانات من داتا الإتصالات وغيرها، مع تحقيق حماية وخصوصية المواطنين؟

إننا نتطلع من خلال هذه الدراسة، من تقديم دراسة حول إمكانية استخدام طرق تُسمّى بالتنقيب عن البيانات وجدواها، مع الإشارة إلى أنّ محور البيانات المستخدمة هي بيانات وهمية تحاكي تلك الموجودة لدى شركات مشغلي الشبكات الخلوية، ولدى الأمن العام اللبناني.

**الكلمات المفتاحية:** تنقيب البيانات، الكشف عن الإرهاب، خوارزميات التصنيف.

## 1- المقدمة

### 1.1- مفهوم الإرهاب

تُشتق كلمة الإرهاب من رهب، رهبًا ورهبة، ووفق تعريفات المجمع اللغوى، فإنّ كلمة الإرهاب ككلمة حديثة فى اللغة العربية أساسها "رهب" بمعنى خاف، لذلك فإنّ تعبير "الإرهابيين" ما هو إلا وصف يطلق على الذين يسلكون سبل العنف لتحقيق أهدافهم العقائدية أو السياسية أو الايديولوجية. تعود الأعمال الإرهابية إلى قديم الزمن لم يستحدث فى تاريخنا المعاصر، إلا أنه لأسباب ربما مرتبطة ببعض التعقيدات السياسية أو الدينية فقد أصبح مفهوم هذه العبارة غامضًا أحيانًا ومختلفًا عليه في أحيان أخرى.

يُعدّ الإرهاب وسيلة من وسائل الإكراه في المجتمع الدولي. تعريف القانون الجنائي للإرهاب تشير إلى تلك الأفعال

العنيفة التي تخلق أجواء من الرعب، وعادة ما يكون موجهًا ضد أتباع دين أو حزب أو هدف أيديولوجي معين، وفيه استهداف متعمد أو تجاهل سلامة غير المدنيين [1].

### 1.2– خصوصية المواطن

الخصوصية هي حقّ الأفراد أو المجموعات أو حتى المؤسسات الاعتبارية في أن تقرر كيفية التعامل ونقل المعلومات الخاصة بها من حيث: متى، كيف، كمية، جهة، وشكل. وهذا يشمل خصوصية المعلومات: المكانية، الشخصية، والمعلومات. كذلك المعلومات التعريفية التي تشمل: الاسم الشخصي، رقم الهوية، الصورة الشخصية، رقم رخصة القيادة، عنوان البريد الالكتروني، العنوان الشخصي، رقم الهاتف النقال، رقم الهاتف، قيمة الراتب، وغيرها من المعلومات التي تميز الأفراد والمؤسسات بعضهم عن بعض. إضافة إلى تلك المعلومات التي تتعلق بتنقلات الفرد داخل وخارج الوطن، ممتلكاته الشخصية، علاقاته الإجتماعية وغيرها من المعلومات. يعتبر الحقّ فى الخصوصية عميق الجذور وهى أحد الحقوق الأساسية التى نصّت عليه الكتب السماوية، ولعله الحقّ الذى يزداد التركيز عليه فى الوقت الحاضر فى ظل إفرازات وأثار توظيف تقنيات المعلوماتية الحديثة [2]

### 1.3– من تحليل للبيانات إلى التنقيب فيها

أدى الانتشار الواسع لتقنية المعلومات وسهولة إتاحتها إلى تضخم حجم المعلومات بصورة استباقية لم يشهدها التاريخ من قبل، مما جعل من قضية البيانات الضخمة على الإنترنت وشركات الإتصالات مثلا مثارًا للجدل، من حيث جدوى وجودها بهذه الصورة العشوائية. وعندما نتحدث عن البيانات الضخمة، فإننا نتحدث عن كميات لا يمكن تخيلها من البيانات متعددة الأنواع والمصادر بحجم يصل إلى المئات من التيرابايت أو حتى البيتابايت (البيتابايت هو الرقم واحد متبوعا بـ15 صفرًا).

من هنا ظهر ما يسمى باستخراج البيانات أو تنقيب البيانات Data Mining كتقنية تهدف إلى استنتاج المعرفة من كميات هائلة من البيانات، تعتمد على الخوارزميات الرياضية والتي تعدّ أساس التنقيب عن البيانات وهي مستمدة من العديد من العلوم مثل علم الإحصاء والرياضيات والمنطق وعلم التعلم، والذكاء الاصطناعي والنظم الخبيرة، وعلم التعرف إلى الأنماط، وعلم الآلة. وغيرها من العلوم، التي تعدّ من العلوم الذكية وغير التقليدية. ظهر التنقيب في البيانات (Data mining) في أواخر الثمانينات وأثبت وجوده كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات، وذلك بتحويلها من مجرد معلومات متراكمة وغير مفهومة (بيانات) إلى معلومات قيّمة يمكن استغلالها والاستفادة منها بعد ذلك. برز آنذاك كمجال حديثٍ ذي قيمة بحثية في دراسة ما يُسمّى بالذكاء الصناعي وقواعد البيانات وتعلم الآلية والإحصائيات وعرض البيانات وغيرها [3].

يُعدّ التنقيب عن البيانات عملية متطورة تقوم باستنتاج البيانات المطلوبة من كم



كبير من البيانات طبقا لأهداف محددة مسبقًا [4].

تنقيب البيانات هو حقل متعدد التخصصات، يستفيد من المجالات بما في ذلك تقنية قاعدة البيانات، الذكاء الاصطناعي، والتعلم الآلي، والشبكات العصبية، والإحصاءات، والتعرف على الأنماط، والنظم القائمة على المعرفة، واكتساب المعرفة، واسترجاع المعلومات، والحوسبة عالية الأداء والصورة ومعالجة الإشارات، وتحليل البيانات المكانية والبيانات التصورية (Data Visualization) التي تعتمد بشكل كليّ على الإدراك البصري.

وقد اجتذبت مرحلة التنقيب في البيانات الكثير من الاهتمام في الأوساط البحثية على مدى العقد الماضي، في محاولة لتطوير خوارزميات قابلة للتوسع والتكيف مع كميات متزايدة من البيانات في البحث عن أنماط معرفية ذات معنى. وقد نمت حزم من الخوارزميات والبرمجيات وبشكل كبير خلال العقد الماضي، إلى حدّ أن التوسع قد جعل من الصعب على العاملين في هذا الحقل تتبع التقنيات المتاحة لحل مهمة معينة.

اكتشاف المعرفة في قواعد البيانات ( Knowledge Discovery in Database KDD) ليس بالعملية السهلة والتي قد يعتقد البعض أنها تتوقف عند تجميع البيانات وإدارتها، بل نراها تمتد إلى التحليل والتوقع والتنبؤ بما سيحدث مستقبلًا.

التنقيب في البيانات يشكل جزءا من اكتشاف المعرفة knowledge discovery، وهذه العملية هي الأكثر شمولا. تتضمن عملية اكتشاف المعرفة الخطوات التالية والتي تتدرج ضمن الشكل1:

**أ) اكتشاف البيانات Data discovery:** وهي مرحلة جمع البيانات وتشمل كشف وتحديد وتوصيف البيانات المتاحة.

**ب) تصفية البيانات وتنقيتها Data cleaning:** ويتم في هذه المرحلة إزالة البيانات المزعجة Noise التي لا أهمية لها، كما يتم حذف البيانات المتضاربة والبيانات غير المتناسقة.

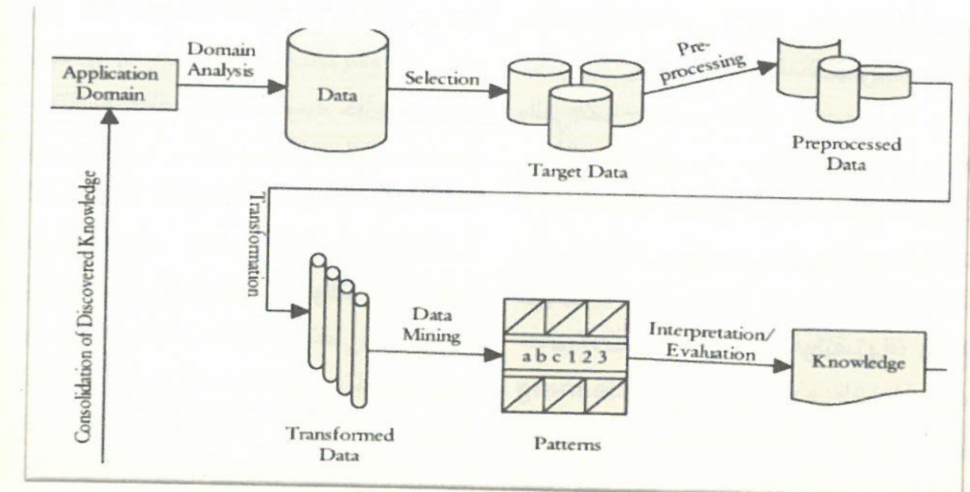
**ت) تكامل البيانات Data integration:** يتم في هذه المرحلة تجميع البيانات المتشابهة وذات الصلة من مصادر البيانات المتعددة ودمجها معا.

**ث) اختيار البيانات Dataselection:** في هذه المرحلة، يتم تحديد واسترجاع البيانات الملائمة من مجموعة البيانات.

**ج) تحويل البيانات Data transformation:** في هذه المرحلة يتم تحويل البيانات إلى نماذج مخصصة ملائمة لإجراءات البحث والاسترجاع بواسطة خلاصة الإنجاز أو عمليات التجميع.

**ح) التنقيب عن البيانات Data mining:** أي استخدام طرق ذكية تطبق لاستخلاص أنماط البيانات، واستخراج نماذج مفيدة قدر الإمكان.

**خ) تقييم النمط Pattern evaluation:** يتم في هذه المرحلة تحديد الأنماط المهمة حقًا والتي تمثل قاعدة المعرفة لاستخدام بعض المقاييس المهمة.



الشكل 1: مراحل اكتشاف المعرفة

د) **تمثيل المعرفة وتقديمها Knowledge presentation:** وهي المرحلة الأخيرة من مراحل اكتشاف المعرفة في قواعد البيانات وهي المرحلة التي يراها المستفيد، هذه المرحلة الأساسية تستخدم الأسلوب المرئي لمساعدة المستفيد في فهم وتفسير نتائج استخراج البيانات.

إذًا التنقيب في البيانات هو خطوة أساسية لتطبيق أساليب ذكية بهدف الكشف عن أنماط البيانات المثيرة للاهتمام والمخبأة في مجموعات البيانات الكبيرة. ومع ذلك، في بعض المنظمات نجد أن مصطلح التنقيب عن البيانات "data mining" أصبح أكثر شعبية للإشارة إلى العملية التي يتم فيها اكتشاف المعرفة knowledge discovery برمتها.

وهناك جانب مهم جدًا، يجب النظر إليه في الاعتبار وهو أن هناك أنماطًا جديدة قد تبرز، عادة ما تكون غير معروفة من قبل. وبالتالي يجب أن تكون أدوات التنقيب عن البيانات قادرة على البحث عن أنواع مختلفة من الأنماط، بأشكال متوازية لزيادة كفاءة التنقيب عن البيانات. كما يجب أيضا أن يتم الكشف عن الأنماط في الأجزاء الصغيرة والفرعية، والتي تعرف بالحبيبات granularities، مما يعني البحث في مستويات مختلفة من التجريد أو التفصيل.

حلول التنقيب عن البيانات الجيدة هي التي تشير أيضًا إلى قدر من الثقة أو اليقين المرتبطة مع نمط اكتشافها، لأن بعض الأنماط قد لا تصلح لكافة البيانات في مجموعة البيانات التي تم تحليلها.

## 2- الدراسات السابقة والمثيلة

في السياق القانوني، تعد أدواة تنقيب البيانات من أهم التقنيات المستخدمة في عمليات البحث أو التحليلات في قواعد البيانات من أجل اكتشاف أو تنبؤ أو حالات شذوذ تدل على وقوع عمل إرهابي أو إجرامي محتمل. هذه الأداة تمنح السلطات الأمنية القدرة على جمع



المعلومات في كثير من الأحيان من مصادر مفتوحة على شبكة الإنترنت أو شبكات الهاتف الخلوي أو الثابت لاستخراج بعض المعرفة التي من شأنها قد تكون مفيدة لهم.

وتهدف تطبيقات استخراج البيانات ضد الإرهاب إلى جمع البيانات الشخصية الموجودة على شبكة الإنترنت أو الهاتف الخلوي أو الثابت، ومن ثم محاولة ربطها ببيانات أخرى من ملفات قضائية أو جرمية أو ما شابه.

نستعرض في ما يلي أعمالا مشابهة لدراستنا هذه من أجل إبراز خصوصية بحثنا فيما يتعلق بتطبيقه في المجال القضائي.

### 2.1- تجميع عملاء شبكات الهاتف المحمول

واحدة من أهم تطبيقات تنقيب البيانات، هي تلك المطبقة من أجل خدمة تسويقية أفضل على بيانات شبكات الهاتف المحمول[5] ، دفعت المنافسة بين مشغلي شبكات الهاتف الخلوي الشركات إلى استخدام أنظمة أوتوماتيكية لتحليل البيانات المسجلة لديها، بإعتبار أنّ سلوك العملاء هو عامل يؤثر في تحسين استراتيجية الشركة، التي تعمل على تقسيم أو تجميع العملاء ضمن فئات اجتماعية من ناحية أخرى.

يعدد المؤلفون الكثير من أساليب التجميع، لكنه يقدم ويشرح بالتفصيل الطريقة الأكثر استخدامًا «K-means». ويشار إلى أن المؤلفين قد طبقوا طرائق تجميع البيانات دون أي تعديل للمتغيرات أو حتى تصفية البيانات. بينما في بحثنا، لقد تغلبنا على أوجه القصور، وشرعنا في تجزئة البيانات وتحليلها عن طريق الاعتماد على تحليل الخبراء.

### 2.2- الكشف عن الأنشطة الإرهابية على شبكة الإنترنت والهاتف الخلوي باستخدام تقنيات تنقيب البيانات

بينت دراسة [6] بوضوح دور تنقيب البيانات في رصد الأنشطة الإلكترونية على شبكة الإنترنت من أجل الكشف عن الأعمال الإرهابية. ووفقًا لهؤلاء الباحثين، فإن الهدف الرئيس هو تحليل سلوك المستخدم النهائي على الشبكة العنكبوتية بما في ذلك تحليل لكافة الصور والفيديوهات المستخدمة والتواقيع الإلكترونية.

وقد طبقت الدراسة [7] تقنيات التنقيب في البيانات على حالة واقعية لمكافحة الإرهاب. لذا، اتخذت شبكة الهاتف المحمول في نيجيريا كبيئة للبحث. وأوضحوا دور استخراج البيانات لمكافحة الإرهاب مباشرة أو في حالة المكالمات المسجلة.

وبالإضافة إلى ذلك، حدد الباحثون دور تقنيات التصنيف والتجميع. كما أنهم أظهروا الدور المهم لتحليل الشبكات الاجتماعية في مكافحة الإرهاب، من خلال تقسيم بسيط للبيانات استنادا إلى شبكات الهواتف المحمولة والشبكات GPS باستخدام خوارزمية (k –means). يشار إلى أن الباحثين في [7] لم يتعاملوا مع المسارات التالية:

- مرحلة تدريب البيانات الوصفية.
- متغيرات القرار.

- دور الخبراء القضائيين.
- تطبيق خوارزميات التصنيف والتجميع.

في سياق مختلف بينت دراسة [8] أن الحكومة لا يمكنها أن تستخدم تقنيات تنقيب البيانات كأداة كافية في مكافحة الإرهاب.

تسلط هذه الدراسة الضوء على عيوب تنقيب البيانات التي تقوض خصوصية المواطنين، وتوضح دورها في عالم التعرف إلى الأنماط، في حين أن الشكل المدروس هو الصوت. وأوضح الباحثون أنه يتم تطبيق هذه التقنيات من قبل مكتب التحقيقات الفدرالي (مكتب التحقيقات الفدرالي) وNSA (وكالة الامن القومي) للبحث عن مشتبه بهم عن طريق تحليل المكالمات الهاتفية بالبحث عن العبارات والكلمات محددة بين ترليونات المكالمات الهاتفية.

### 2.3- الكشف عن تبيض الأموال باستخدام تنقيب تنقيب البيانات

وصف الباحثون في دراستهم [9]، دور تنقيب البيانات في الكشف عن عمليات تبييض الأموال، من دون التطرق إلى شرح تقنيات تحليل المعاملات المالية والمصرفية. وبيّن الباحثون كيف أنّ طرائق "التعدين أو التنقيب المتكرر للتسلسل" إذا ما كانت قد تمثل عمليات تبييض الأموال أم لا.

وسمّى الباحثون طرائق من تقنيات تنقيب البيانات وكيفية استخدامها للهدف المنشود، وهذا ما هو مبين في الجدول التالي:

| التقنية | الهدف |
|---|---|
| قواعد الجمعيات (Association rules) | الكشف عن العلاقة الخفية بين المعاملات المالية والمشاركين |
| التنقيب في التسلسل المتكرر ( frequent (sequence mining | كشف أنماط المعاملات التي تحدث في كثير من الأحيان |
| خوارزميات التصنيف Classification ) (algorithms | تساعد على تصنيف الحسابات إلى فئات محددة سلفا من المخاطر، تبعا لملامح المخاطر وأصحاب الحسابات |
| خوارزميات التجميع Clustering ) (algorithms | تجميع المعاملات / الحسابات إلى مجموعات من المعاملات / الحسابات المماثلة على أساس أوجه التشابه. تساعد خوارزميات التجميع في بناء التشكيلات الجانبية للتسلسل المشبوه للمعاملات ومنهم في تحديد خصائص مخاطر العملاء / الحسابات. |
| تحليل الإنحدار ( Regression (analysis | تتوقع إمكانية استخدام حساب كقناة لغسل الأموال على أساس المتغيرات الديموغرافية والسلوكية. |
| وصلة التعدين والتحليل Link mining and ) (analysis | يبرز اتصالات خفية بين حسابات مختلفة على أساس معايير مثل نشاط تحويل الأموال والتفاعل مع نفس الحسابات أو ما شابه ذلك. |

**جدول1: دور تقنيات تنقيب البيانات للكشف عن عمليات تبيض الأموال حسب [9]**

### 2.4- تحليل ملامح الجريمة باستخدام تقنيات تنقيب البيانات

تطبيق تقنيات تنقيب البيانات من أجل تحليل ملامح الجريمة هي محور دراسة [10]. حيث شرح الباحثون في أهمية مراحل معالجة البيانات وتنظيفها، فضلا عن نتائج تحليلات البيانات التي يمكن تقييمها لتقييم معرفة ممكنة. لضمان الهدف المرجو من الدراسة، يوضح الباحثون كيفية استخراج البيانات لتحليل ملامح الجريمة



من أجل تحديد الاتجاهات المهنية الجنائية.

### 2.5- المراقبة بالفيديو الذكي

أبرز الباحثون[11] دور النظام الذكي في عمليات مراقبة الفيديو باعتماد ما سموه نظام الفيديو الذكي. يعتمد فيديو المراقبة الذكي على أنظمة تلقائية مثل معالجة الإشارات والذكاء الاصطناعي واستخراج البيانات لتوسيع نطاق ميزات المراقبة بالفيديو والتطبيقات. إلى:

1. توقع الحوادث عن طريق الكشف عن السلوك المشبوهة وإطلاق الإنذارات في الوقت الحقيقي.

2. مساعدة عمليات التحقيق وتحسينها من خلال إجراء بحث في المحتوى، والمتابعة المكانية والزمانية.

وبطبيعة الحال، فإن نظام المراقبة بالفيديو هذا غير متاح للعملية نظرا لتكلفة البنية التحتية (الكاميرا والشبكات والخوادم وما إلى ذلك) وتعقيد المعالجة المطلوبة.

### 3- هدف الدراسة

إننا نهدف من هذه الدراسة العلمية الى تقديمها بين يدي السلطات الأمنية والقضائية اللبنانية لبيان كيفية الإستفادة من الأنظمة والتقنيات الأوتوماتيكية لإستخلاص المعلومات وتحليلها ذلك لتسهيل التحقيق والكشف عن مشبوهين في عمليات إرهابية في البلاد، يكمن التحدي في هذه الدراسة الى الوصول الى أشخاص مشتبه بهم مع تقليل نسبة الوقوع في التشخيص الخطأ، مع الأخذ في الإعتبار خصوصية المواطنين واحترامها.

### 4- المنهجية المتبعة ومصدر البيانات

لأننا نعلم بأنّ الإشكالية المتناولة دقيقة جدًا، ولأنّ حجم البيانات كلما زاد، زادت كمية المعرفة المتوقعة، ولأنّ زيادة حجم البيانات بشكل عشوائي يؤدي أيضًا إلى التشخيص الخاطئ، عمدنا إلى انشاء استمارة استبيانية، تقودنا الى استنتاج ماهية العوامل والمتغيرات المهمة التي تقود عادة المحققين الى "طرف الخيط" في تحقيقاتهم.

الاستمارة المذكورة ملأها مجموعة كبيرة من قضاة ومحامين ومحققين ومساعدين قضائيين وغيرهم من الأشخاص المعنيين في السلك الأمني والقضائي.

التحليل الإحصائي للإستمارات خلص إلى تسمية عدد من المتغيرات المساعدة في عملية استخلاص البيانات المخطط لها.

نذكر من هذه المتغيرات:

- عدد مرات الدخول والخروج من البلاد.
- وجود سوابق جرمية لمستخدمي الخطوط الخلوية.
- الأماكن التي تم منها شراء الخط الخلوي.
- حركة استخدام الخطوط الخلوية بعد حصول الجريمة.
- التاريخ الذي تم فيه شراء خطوط خلوية.
- اقفال الخط الخلوي بعد تاريخ ارتكاب الجريمة.

– استخدام الخطوط الخلوية من قبل أشخاص وفدوا إلى لبنان قبل حصول الجريمة.

بات الان واضحاً بأنّ مصدر البيانات حسب المتغيرات المذكورة هو:

– الشركات المشغلة لشبكات الاتصالات الخلوية في لبنان.

– الأمن العام اللبناني.

إن الحصول على بيانات من المصادر المذكورة مهمة شبه مستحيلة، وإن كانت من أجل البحث العلمي. لذلك، عمدنا الى انتاج بيانات وهمية تحاكي واقع البيانات الحالي وذلك بعد حصول جريمة معينة، بهدف بناء الدراسة عليها.

**5- تحضير البيانات Data preparation:**

تعدّ هذه المرحلة مرحلة تمهيدية لتحليل البيانات، وتسمّى هذه المرحلة مرحلة المعالجة التمهيدي Pre-processing أو مرحلة تنظيف البيانات Data cleaning.

تهدف هذه المرحلة الى تحضير البيانات للمعالجة، خصوصا أننا ذكرنا أن البيانات المخطط لدراستها تم انتاجها وهميًا وذلك لتعذر الحصول عليها.

عدة طرق وتقنيات استخدمت في هذه المرحلة: جدولة البيانات وازالة النواقص والاخطاء، وفحص جودتها ونزع غير الملائم منها أو تصحيحه. لذلك، قمنا بإعطاء قيم جديدة لكل متغير، فعلى سبيل المثال، تصنيف بيانات المتغير "حركة استخدام الخطوط الخلوية بعد حصول الجريمة" إلى فئات ثلاث:

الفئة الأولى: (القيمة =1): خارج البلاد.

الفئة الثانية: (القيمة =2): منطقة تأوي ارهابيين – معرفة مسبقًا.

الفئة الثالثة: (القيمة=3): منطقة داخل البلاد غير مشبوهة.

**6- تجميع البيانات Data clustering**

تجميع البيانات هي عملية وضع البيانات في تجمعات متشابهة. تسعى هذه الطريقة الى تصنيف البيانات الى كتل متشابهة في خصائصها.

تقسم خوارزمية التجميع مجموعة بيانات الى عدة تجمعات، حيث انّ التشابه والتقارب بين نقطتين ضمن تجمع معين أكبر من تشابه بين نقطتين في تجمعين منفصلين. بالنسبة لنا في الدراسة هذه كل نقطة هي عبارة عن شخص ما، لدينا مسبقا بيانات عنه.

تعد خوارزمية K-means clustering من أبرز الخوارزميات المستخدمة في تجميع البيانات، لذلك استخدمناها في هذا البحث مع تحديد (K=2) عدد المجموعات التي نود أن نقسم البيانات اليها الى مجموعتين اثنين.

وبعد تطبيق هذه الخوارزمية، خلص الينا مجموعتين من البيانات التي عرضناها على خبراء في مجال التحقيق الأمني، فأكدوا لنا أنه من الواضح جدًا أن التقسيم بدا بأن المجموعة الأولى التي تحوي حوالي 87% من البيانات، تعد المجموعة التي تمثل بيانات الاشخاص غير المشبوهين في حين أنّ الأخرى تمثل الأشخاص المشبوهين.



لذلك، أضفنا متغيرًا جديدًا يمثل إذا كانت البيانات مصنفة ضمن المجموعة الأولى أم المجموعة الثانية.

**7- خوارزميات التصنيف Data classification**

في المرحلة السابقة، أضفنا متغيرًا جديدًا يمثل ما يُسمى بـ"class" لكل شخص، وعليه يصنف ما اذا كان هذا الشخص مشبوهًا أم لا.

هنا في هذه المرحلة، نعرّف ما يسمّى بـ"خوارزميات التصنيف" أي الخوارزميات التي من شأنها أنْ تحدد الإنتماء للشخص لأي مجوعة من المجموعتين ينتمي. ومن أجل تحقيق الهدف المرجو، قمنا بتطبيق عدة خوارزميات على البيانات التي بين أيدينا.

**7.1- التصنيف وفق خوارزمية naïve Bayesian**

تعتمد خوارزمية naïve Bayesian على مبرهنة Bayes والتي تستند على الاحتمالات المشروطة. فهي صيغة تحتسب احتمالية الإنتماء لكل صنف موجود.

أخذت المبرهنة هذا الاسم نسبة إلى توماس بايز الذي توصل الى النتائج الأولية التى أستخدمت فيما بعد للحصول على المبرهنة بشكلها النهائي، فقد استخرج الرياضى الفرنسى لابلاس المعادلات المبنية على أساس الاحتمالات وهو الشكل النهائى الذى انتشرت فيه هذه المبرهنة بعد أن قام بايز بكتابتها بالتكاملات.

المعادلة الرياضية التي ينبني عليها هذا المصنف هي كما تسمّى في مجال الاحتمالات معادلة (Bayes):

حيث أنه يمكن أن يكون:

خصائص = خاصية 1 + خاصية 2 +...+ خاصية n

وحساب احتمال الخصائص "خ" علمًا الصنف "ص" هو كالتالي:

احتمال الخصائص "خ" = احتمال الخاصية 1 x احتمال الخاصية 2 x...x احتمال الخاصية

وحساب احتمال الخصائص "خ" علمًا الصنف "ص" هو كالتالي:

احتمال الخصائص "خ" علمًا الصنف "ص" = احتمال الخاصية 1 علمًا الصنف "ص" x احتمال الخاصية 2 علمًا الصنف "ص" x...x احتمال الخاصية علمًا الصنف "ص".

يشار إلى أنّه لتطبيق هذه الخوارزمية يفرض بالمتغيرات أن تكون مستقلة ولا يوجد ارتباط فيما بينها.

احتمال الصنف "ص" علما الخصائص "خ" = (احتمال الخصائص "خ" علما الصنف "ص" x احتمال الصنف "ص") / (احتمال الخصائص "خ")

## 7.2- التصنيف وفق خوارزمية Bayesian Network

تعتمد Bayesian Networks على نموذج رياضي للاستدلال الاحتمالي، ويتم الاستدلال الاحتمالي من خلال بعض المعلومات للحصول على احتمالات للمتغيرات الأخرى، وتعتمد شبكات النظرية الإفتراضية على أساس الاستدلال الاحتمالى لحل مشكلة عدم اليقين [12]. فهى مخططات موجهة غير حلقية مؤلفة من مجموعة عقد تمثل متغيرات مختلفة ومجموعة أقواس تمثل العلاقات الاعتمادية (غير المستقلة dependence relation) بين هذه المتغيرات.

إذا كان هناك قوس يتجه من العقدة A إلى العقدة B، عندئذ يمكن أن نقول أن العقدة A هي والد أو أصل العقدة B. إذا كانت للعقدة قيمة معروفة (ثابتة) عندئذ تدعى (عقدة تأكيد node evidence) يمكن للعقد أن تمثل أي نوع من أنواع المتغيرات: قياسات، مؤشرات (معالم parameter)، أو فرضيات hypothesis.

تدعى أيضًا شبكات الاعتقاد البايزي Bayesian belief network أو اختصارا شبكات الاعتقاد belief network ولها تطبيقات عديدة في حقل المعلوماتية الحيوية.

تمثل الشبكات البايزية التوزيع الاقتراني للمتغيرات كافة الممثلة بعقد الشبكة. إذا افترضنا المتغيرات التالية: X(1), ..., X (n)، وليكن مصطلح أصول (A) التعبير عن مجموعة العقد المتصلة بالعقدة A عندئذ يكون التوزيع الاقتراني للمتغيرات من $X(1)$ إلى $X(n)$ مثل جداء التوزيعات الاحتمالية:

$$\Pr(X(i) \mid \text{parents}(X(i))$$

من أجل: iالذي يأخذ قيمًا من 1 إلى n.

إذا لم تكن للعقدة والد (أصل) عندئذ يكون توزعها الاحتمالي **غير شرطي unconditional**، وإلا فإن توزعها الاحتمالي يدعى شرطي (عندما يكون لها والد).

## 7.3- التصنيف وفق خوارزميات Lazy

المتعلمون الكسولون Lazy learners هم حالات من تخزين التدريب والقيام بأي عمل حقيقي حتى وقت التصنيف. التعلم الكسول هو أسلوب الذي يتأخر في تعميم البيانات التدريبية إلى أن يتم الاستعلام إلى النظام حيث يحاول النظام تعميم بيانات التدريب قبل تلقي الاستفسارات.

والميزة الرئيسة المكتسبة في توظيف طريقة التعلم الكسول هي أن وظيفة الهدف الذي سيتم تقريبه محليا مثل في خوارزمية k-means. ولأن الدالة الهدف تقترب محليًّا لكل استعلام للنظام، فإن أنظمة التعلم الكسولة يمكن أن تحلّ في وقت واحد مشاكل متعددة وتتعامل بنجاح مع التغيرات في ساحة المشكلة [13].

من أهم مساوئ تعلم الكسول أنها تشمل متطلبات مساحة كبيرة لتخزين مجموعة كاملة من التدريب. وتزيد بيانات التدريب الصاخبة في معظم الأحيان من دعم الحالة



دون داع، لأنه لا يوجد مفهوم أثناء مرحلة التدريب، وهناك عيب آخر هو أن أساليب التعلم كسول عادة ما تكون أبطأ للتقييم، على الرغم من أن هذا يرتبط مع مرحلة التدريب أسرع.

### 7.3.1- التصنيف وفق خوارزمية IBK

IBK هو المصنف k-means الذي يستخدم مقياس المسافة نفسها. يمكن تحديد عدد أقرب الجيران بشكل صريح في محرر الكائن أو تحديده تلقائيًا باستخدام التركيز البيني للمصادقة عبر الإجازة إلى حد أعلى تعطى القيمة المحددة. إيبك هو مصنف قريب الجوار k. وهناك نوع من خوارزميات البحث المختلفة يمكن استخدامها لتسريع مهمة العثور على أقرب الجيران.

ودالة المسافة المستخدمة هي معلمة لطريقة البحث. الشيء المتبقي هو نفسه ل إبل- وهذا هو، المسافة الإقليدية. وتشمل الخيارات الأخرى بحسب [14] أنه يمكن ترجيح التوقعات من أكثر من جار واحد وفقًا لمسافاتها عن مثيل الاختبار، ويتم تنفيذ صيغتين مختلفتين لتحويل المسافة إلى وزن [13].

يمكن تقييد عدد حالات التدريب التي يحتفظ بها المصنف عن طريق تحديد خيار حجم النافذة. كما يتم إضافة حالات التدريب الجديدة، أقدمها منفصلة للحفاظ على عدد من حالات التدريب في هذا الحجم.

### 7.3.2- التصنيف وفق خوارزمية Kstar

يمكن تعريف خوارزمية كستار على أنها طريقة لتحليل المجموعات التي تهدف أساسا إلى تقسيم الملاحظة n إلى مجموعات k حيث تنتمي كل ملاحظة إلى المجموعة بأقرب متوسط. يمكننا أن نصف خوارزمية K* كمتعلم القائم على المثال الذي يستخدم الإنتروبيا كمقياس المسافة. من فوائده أنه يوفر نهجًا متسقًا للتعامل مع الصفات القيمة الحقيقية، والسمات الرمزية والقيم الناقصة [15]. ومن هنا يمكن ان نعرف K* هو بسيط، مثبت القائم على المثال، على غرار K- أقرب الجار (K-ن). يتم تعيين مثيلات بيانات جديدة، x، إلى الفصل الذي يحدث بشكل متكرر بين نقاط البيانات الأقرب إلى k. ثم يتم استخدام المسافة إنتروبيك لاسترداد الحالات الأكثر مماثلة من مجموعة البيانات. من خلال المسافة الإنتروبية كمقياس له عدد من الفوائد بما في ذلك التعامل مع الصفات القيمة الحقيقية والقيم المفقودة.

## 7.4- خوارزميات التصنيف وفق قواعد (Rule classifier algorithms)

### 7.4.1- التصنيف وفق خوارزمية One R

تعد خوارزمية One R واحدة من أبسط خوارزميات التصنيف. كما هو موضح في [16]، تنتج قواعد بسيطة تستند إلى سمة واحدة فقط. فإنه يولد شجرة القرار على مستوى واحد، والتي يتم التعبير عنها بمجموعة من القواعد لكل اختبار سمة معينة واحدة.

إنها طريقة بسيطة غالبا ما تأتي مع قواعد جيدة جدًا لتوصيف الهيكل في البيانات [17]. غالبًا ما يحصل على دقة

معقولة على العديد من المهام ببساطة عن طريق النظر في سمة واحدة.

لكل سمة: A

لكل قيمة V من تلك السمة، قم بإنشاء قاعدة:

1- عد عدد المرات التي تظهر فيها كل فئة

2- العثور على التصنيف class الأكثر شيوعا، c

3- جعل قاعدة "إذا A = V ثم C = c"

احسب معدل الخطأ لهذه القاعدة واختر السمة التي تنتج قواعدها أدنى معدل خطأ.

### 7.4.2- التصنيف وفق خوارزمية Zero R

هي الطريقة الأبسط للتصنيف وتعتمد على الهدف مع تجاهل كل التوقعات المسبقة. هي طريقة تصنيف تتوقع فئة الأكثرية. هي طريقة مفيدة لتحديد نقطة مقارنة لمختلف طرق التصنيف. تعتمد الخوارزمية على إنشاء جدول تردد للهدف وتحديد القيمة الأكثر شيوعًا. التنبؤات المساهمة: لا يوجد شيء يمكن أن يقال عن مساهمة التنبؤ إلى نموذج لأن زيرور لا تستخدم أي منها. تقييم نموذج: زيرور يتنبأ فقط فئة الأغلبية بشكل صحيح. وكما ذكر من قبل، فإن زيرور مفيد فقط لتحديد أداء خط الأساس لطرق التصنيف الأخرى [18].

### 7.4.3- التصنيف وفق خوارزمية LWL (Locally weighted learning)

تعتمد خوارزمية LWL إلى تسنيد أوزان مدروسة لكل مثال أو حالة وفق ما يُسمى بـ«weighted instances handler» [19]. ثم بعد ذلك يتم تفعيل التصنيفات بإستخدام Bayesian Networks أو الإنحدارت Regression على سبيل المثال.

### 7.4.4- التصنيف وفق خوارزمية Ridor

تقوم خوارزمية Ridor بإنشاء قاعدة افتراضية أولًا، ثم الاستثناءات للقاعدة الافتراضية مع معدل الخطأ (المرجح) الأقل. ثم يولد "أفضل" استثناءات لكل استثناء، ويكرر حتى تقليص معدل الخطأ. وبالتالي فإنه يؤدي توسع تشبه شجرة من الاستثناءات. استثناءات هي مجموعة من القواعد التي تتنبأ التصنيفات الأخرى Classes بخلاف الافتراضي أو ما يسمى بال default [20]. تعتبر هذه الخوارزمية كنهج تدريجي في اكتساب المعرفة.

## 8- النتائج

الإختبار كان بأخذ البيانات السابقة، وأبقينا منها 66% منها كما هي، وذلك من أجل أن تكون قاعدة انطلاق وقياس لهذه الخوارزميات، في حين أنّ البيانات المتبقية استخدمت لإختبار الخوارزميات المذكورة من أجل اعادة تصنيفها ومقارنتها بالأصل إذا ما كان التصنيف صحيحًا أم لا. وقد حصلنا على أداء الخوارزميات حسب الجدول 2، ويبين الشكل 2 الوقت المستغرق لكل خوارزمية. يبدو واضحا أنّ أداء LWL وOneR هما الأفضل من حيث دقة التصنيف (100%) ويظهر جليًا أن LWL هو الأفضل من بين الخوارزميات من حيث الوقت المستغرق للتصنيف.



لقد بينّا في هذه الدراسة أننا يمكننا أن نصل إلى دقة في تحديد إرهابيين محتملين تصل إلى 100%، وهذا يعني أنه إذا أعطت الدراسة نتائج حول إرهابيين محتملين فإحتمال الخطأ بأن يكونوا ليسوا كذلك هو 0%. إننا نتطلع لتنفيذ هذه الدراسة على بيانات حقيقية، نقدم من خلالها دراسة المعنيين بجريمة ما لتقديمها بين يدي السلطات.

| الخوارزمية | نسبة التصنيف الخاطئ | نسبة التصنيف الصائب |
|---|---|---|
| naïve Bayesian | 9.2437 | 90.7563 |
| Bayesian Network | 0.4202 | 99.5798 |
| IBK | 20.1681 | 79.8319 |
| Kstar | 0.4202 | 99.5798 |
| ZeroR | 30.6723 | 69.3277 |
| LWL | 0 | 100 |
| OneR | 0 | 100 |
| Ridor | 0.8403 | 99.1597 |

جدول 1 : جودة النتائج لكل خوارزمية.




الشكل 1: الوقت المستغرق لكل خوارزمية.

## 9- الخلاصة والتوصيات

إننا من خلال هذه الدراسة، نحاول أن نقدم إلى الأجهزة الأمنية والقضائية شيئًا بسيطًا مما يمكننا فعله من أجل محاربة الإرهاب واجتثاث العقول المدبرة. انّ هذا الطرح الذي خَلُص إلى إمكانية تصنيف الأشخاص كضالعين في جريمة إرهابية، محددة مسبقًا، بـ 0% خطأ ولو على بيانات وهمية، يمكن استخدامه على بيانات واقعية مصدرها الشركات المشغلة للهاتف الخلوي في لبنان والأمن العام اللبناني.

***

## الهوامش

* دكتور ومهندس - أستاذ متفرغ في كلية الاقتصاد وإدارة الأعمال في الجامعة الإسلامية في لبنان
البريد الإلكتروني: Kassem.danach@iul.edu.lb

## - المصادر والمراجع

[1] T. Deen; "Politics: UN member states struggle to define terrorism،" IPS News. Net, المجلد 25، 2005.

[2] بوحجة وكريمة، "حماية الخصوصية المعلوماتية في العصر الرقمي،" 2013.

[3] J. Han ،J. Pei و M. Kamber ،Data mining: concepts and techniques, Elsevier, 2011.

[4] J. Durkin و ,C. Jingfeng ،"Decision Tree Technology And Its Current Research Direction, "Control Engineering ،2005.

[5] Q. Lin, "Mobile customer clustering analysis based on call detail records", Communications of the IIMA ،4، المجلد 7، رقم 4، p. 95 ،2007

[6] Y. Elovici, A. Kandel, M. Last ،B. Shapira و O. Zaafrany ،"Using data mining techniques for detecting terror-related activities on the web", Journal of Information Warfare ،1، المجلد 3، رقم 1، pp. 17-29, 2004 .

[7] R. O. Okonkwo و F. O. Enem, "Combating crime and terrorism using data mining techniques, 10 تأليف "th International conference

IT people centred development, Nigeria Computer Society, Nigeria ،2011 .
[8] B. D. Kreykes, "Data mining and counter-terrorism: the use of telephone records as an investigatory tool in the war on terror،" ISJLP ، المجلد 4، p. 431 ،2008 .
[9] J. S. Zdanowicz, "Detecting money laundering and terrorist financing via data mining", Communications of the ACM, المجلد 47، رقم 5 pp. 53-55 ،2004 .
[10] R. Krishnamurthy و J. S. Kumar, "Survey of data mining techniques on crime data analysis", International Journal of Data Mining Techniques and Applications، المجلد 1، رقم 2، pp. 117-120 ،2012 .
[11] N. Baaziz", La vidéo surveillance automatique: sécurisation du contenu et traitements coopératifs", Université du Québec en Outaouais ،Outaouais ،2007.
[12] D. Heckerman, "Bayesian networks for data mining", Data mining and knowledge discovery، المجلد 1، رقم 1، pp. 79-119 ،1997 .
[13] S. Vijayarani و S. Sudha, "Comparative analy sis of classification function techniques for heart disease prediction", International Journal of Innovative Research in Computer and Communication Engineering، المجلد 1، رقم 3، pp. 735-741 ،2013.
[14] S. Ghosh ،S. Roy و S. Bandyopadhyay ،"A tutorial review on Text Mining Algorithms،" International Journal of Advanced Research in Computer and Communication Engineering المجلد 1، رقم 4، p. 7 ،2012.
[15] T. C. Sharma و M. Jain ،"WEKA approach for comparative study of classification algorithm،" International Journal of Advanced Research in Computer and Communication Engineering, المجلد 2، رقم 4، pp. 1925-1931, 2013 .
[16] R. C. Holte ،"Very simple classification rules perform well on most commonly used datasets،" Machine learning، المجلد 11، رقم 1، pp. 63-90, 1993.
[17] D. H. Wolpert ،W. G. Macready و others, "No free lunch theorems for search", 1995.
[18] chem-eng.utoronto.ca ،"ZeroR،" 12 AUGUST 2016 .[متصل] .
[19] A. Skoglund و M. L. Course ،"Locally weighted learning for control", Artificial Intelligence Review، المجلد 11، pp. 11-73, 1997 .
[20] V. Veeralakshmi و D. Ramyachitra ، "Ripple down rule learner (ridor) classifier for iris dataset،" Issues، المجلد 1، رقم 1، pp. 79-85 ، 2015.
[21] T. Hill و P. Lewicki, STATISTICS Methods and Applications. StatSoft, Tulsa, USA ،2007 .
[22] I. H. Witten و E. Frank ،Data Mining: Practical machine learning tools and techniques ،Morgan Kaufmann ،2005 .
[23] H. Kaushik و R. B. Gajjar, "Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA،" Indian Journal of Research (PARIPEX) Volume المجلد 2.
[24] F. L. Seixas ،B. Zadrozny ،J. Laks, A. Conci و D. C. M. Saade ،"A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer 's disease and mild cognitive impairment", Computers in biology and medicine، المجلد 51، pp. 140-158 ،2014.
[25] M. F. bin Othman و T. M. S. Yau, "Comparison of different classification techniques using WEKA for breast cancer", تأليف 3 rd Kuala Lumpur International Conference on Biomedical Engineering 2006, 2007.

***